Knowledge Discovery using clustering techniques

Ján Paralič
Department of Cybernetics and Artificial Intelligence
Technical University of Košice
Letná 9, 042 00 Košice
Slovak Republic
paralic@tuke.sk

Andreas Rauber
Department of Software Technology
Vienna University of Technology
Resselgasse 3/188, A-1040 Wien
Austria
andi@ifs.tuwien.ac.at

Abstract – The main goal of this paper is to provide results and findings of some preliminary tests with two different clustering techniques suitable for knowledge discovery in databases (KDD). We used Bayesian unsupervised classification (based on the AutoClass system) and the Self-Organizing Map (SOM) as a prominent unsupervised neural network to analyze a real-world database of ticket sales for castles in South Bohemia. These tests should serve as a basis for decisions related to the implementation of the KDD package within the European Union research project GOAL1.

I. INTRODUCTION

The goal of the GOAL project is to develop a generic framework that solves the general issues of Geographic Information Systems (GIS) and Data Warehouse (DWH) interoperability, including DWH feeding, knowledge extraction, interpretation, and security concepts. The feasibility of this framework will be tested on 2 very different real world applications from the GIS domain using environmental sensor data and cultural data, allowing a real world evaluation of the framework.

Regarding the knowledge extraction phase, a KDD (Knowledge Discovery in Databases) package will be developed within this project. The basic idea here is to provide as a core of a number of different algorithms and their combinations. Algorithms able to discover association rules, decision trees and clusters are planned to be integrated.

This paper summarizes results of applying two different clustering mechanisms to one of the target application arenas. The database contains sales data for entrance tickets for tourists to various castles in South Bohemia collected over a three months period (about 25,000 tickets). The goal of these preliminary analyses is mainly to get an overview of the data available and its characteristics in order to later on be able to decide (a) which algorithms offer themselves for which type of analysis tasks and (b) which preprocessing of existing data or collection of additional data is necessary to achieve the goals addressed in the GOAL project.

For our initial experiments we selected two different approaches to cluster analysis. The first is a Bayesian unsupervised classification using the AutoClass system [6]. The main idea is to extract a classification system and probabilistic class assignments for the available data.

As a second system we selected the self-organizing map (SOM) [2], an unsupervised neural network which provides a mapping from high-dimensional feature spaces onto a two-dimensional plain such that similar data are mapped close to each other. This allows for a very intuitive cluster representation and analysis. We furthermore combine the basic SOM algorithm with the LabelSOM method to automatically extract classification from the trained SOM.

The remainder of the paper is organized as follows. Section II presents the data available for the application. Section III then introduces the basic principles of the self-organizing map, with Section IV presenting initial results achieved using this approach. Similarly, Section V presents the AutoClass approach to unsupervised classification, followed again by a presentation of results in Section VI. In Section VII we try to compare the results acquired from both with some lessons learned and an outlook on the next steps given in Section VIII.

II CASTLE ADMISSIONS TICKETS DATA

The cultural application within the GOAL project deals with castles in Southern Bohemia. The goal is to analyze visitor numbers and distributions across the participating castles in order to be able to better plan and organize itineraries and additional attractions covering some or all of the castles involved. The first steps in that direction aim at analyzing the current state of visitor numbers and characteristics, such as the distribution of tour groups, group sizes and high season periods or periods of day.

The basis for these preliminary analyses is formed by the sales data recorded at the castles and booking agencies. The data records sales of admission tickets for visitors to a number of castles in South Bohemia. For our initial experiments we use a subset of the complete data set that was recorded over a period of 3 months and contains about 25000 records for individual tickets.

For each ticket the following information is provided within the database: nationality of the visitors (Czech, German and other Foreigners), total price of the ticket (real number), total number of visitors listed on one group ticket (integer number), number of children listed on one group ticket (integer number), price for services (real number), date of the visit, time of the visit (integer number), order - information if the ticket was ordered in advance (binary value).

Due to the different requirements of the two approaches presented in this paper, the data had to be preprocessed in different ways. For example, with the self-organizing map requiring numerical vector representations of the attributes, the nationality attribute had to be spilt up in 3 separate attributes, each containing binary information on the nationality of the visitors listed on a specific ticket. Other preprocessing steps included the normalization of the individual attributes to value ranges between 0 and 1.

1 This work is supported by the European Union under INCO Grant No. 977091 - GOAL - Geographic Information On-Line Analysis (GIS – Data Warehouse Integration)
The resulting data was used as input to the two algorithms in order to obtain information on the clusters present in the tickets sales data. Based on this information we hope to obtain a clearer understanding concerning the requirements of the next analysis steps within the GOAL project with respect to geographical information contained in the data in terms of castle location.

III. SOM - ARCHITECTURE, TRAINING AND INTERPRETATION

The self-organizing map (SOM) [2] is a popular neural network model adhering to the unsupervised learning paradigm and is being widely used for cluster analysis of high-dimensional input data. In a nutshell, its architecture and training procedure can be described as follows: It consists of a usually 2-dimensional grid of units with \( n \) -dimensional weight vectors, with \( n \) being the dimensionality of the data space. During the training process input signals \( x \in \mathbb{R}^n, x \in \mathbb{R}^2 \) are presented in random order. An activation function, e.g. the Euclidean distance, is used to determine the winning unit as the unit showing the highest activation. Next, the weight vectors of the winner and its neighboring units as identified by a time-decreasing neighborhood function \( \varepsilon \) are modified to represent the various input signals more closely. The result of this training process is a map providing a topology-preserving mapping of the high-dimensional input space to the 2-dimensional map space. The individual weight vectors serve as prototypes for the set of input vectors mapped onto the respective units. Input signals that are located on the same or neighboring units are more similar to each other than signals mapped far apart providing an intuitive representation of the clusters found in the data.

While the data representation itself is quite intuitive, analyzing the characteristics and extent of the clusters proves to be somewhat more difficult. A number of extensions to the standard SOM representation, such as the U-Matrix [5] or the Adaptive Coordinates [3] have been proposed to allow for easier cluster identification. Recently, we developed the LabelSOM [4] method to extract the features that characterize a cluster, allowing to automatically label the map including information on the significance of a certain label and the strength of presence of the according attribute. Due to space considerations we refer to [4] for a more detailed presentation of the SOM and LabelSOM methods. Combining these methods allows you to gain a good overview and basis for interpretation of a trained self-organizing map and the mapped data.

IV. SOM EXPERIMENTAL RESULTS

In this section we present a brief overview of the results obtained by using a \( 5 \times 5 \) SOM to cluster the data points. A description of the SOM is provided in 1, where the 25 units are represented by a \( 5 \times 5 \) grid of squares, onto which the data points are classified. Each unit is labeled with the attributes considered most important to describe the tickets mapped onto the respective unit. These attributes were automatically extracted from the map using the LabelSOM method. Since the price turned out to be the most dominant attribute in the tickets classification showing as the top-ranking label for almost all units, we used it to color the units, with higher-priced areas being indicated by darker shadings. We furthermore provide for each unit the approximate number of data points mapped onto each of the units.

<table>
<thead>
<tr>
<th>Middle groups Czech not ordered (0)</th>
<th>Middle groups Czech not ordered (3)</th>
<th>Middle groups Czech not ordered (1900)</th>
<th>Middle groups Czech not ordered (1250)</th>
<th>Large groups Czech ordered (625)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smaller groups Czech not ordered (1275)</td>
<td>Middle groups Czech not ordered (2600)</td>
<td>Middle groups Czech not ordered (2175)</td>
<td>Middle groups a few children Czech (0)</td>
<td>Large groups a few children Czech (475)</td>
</tr>
<tr>
<td>Middle groups German not ordered (1)</td>
<td>Middle groups German not ordered (175)</td>
<td>Large groups a few children Czech not ordered (less) (2100)</td>
<td>Middle groups many children Czech not ordered (975)</td>
<td>Larger groups some children Czech not ordered (1500)</td>
</tr>
<tr>
<td>Small groups Foreigner not ordered (325)</td>
<td>Large groups a few children Foreigner (750)</td>
<td>Large groups with more children Czech (2125)</td>
<td>Larger groups many children Czech not ordered (less) (700)</td>
<td>Middle groups many children Czech, German not ordered (less) (1125)</td>
</tr>
<tr>
<td>Middle groups Foreigner not ordered (less) (875)</td>
<td>Large groups Foreigner (500)</td>
<td>Large groups Foreigner ordered (little) (525)</td>
<td>Larger groups many children Foreigner ordered (little) (500)</td>
<td>Large groups a few children Foreigner not ordered (less) (250)</td>
</tr>
</tbody>
</table>

Fig. 1: 5 x 5 SOM clustering tickets into approx. 13 clusters, with different shading indicating ticket price.
Based on this SOM representation we can identify about 13 clusters of different size, discerned by ticket price, number of adults, number of children, nationality and order. In the upper left area of the map we can find a large cluster of rather low-priced tickets of middle-sized groups of adults of Czech nationality and "not ordered". The clusters in the upper right area contain the higher-priced group tickets of larger groups of again predominantly Czech nationality. Tickets of international groups can be found in the lower left area of the map, with the tickets for about 500 large groups obviously yielding the highest prices. Groups with larger number of children can be found in the lower right area of the resulting SOM.

![SOM map](image1)

![SOM map](image2)

![SOM map](image3)

![SOM map](image4)

**Fig. 2: Analysis of selected attribute distributions on the SOM.**

The distribution of the various attributes can also be analyzed separately as indicated in 2a - d. Fig. 2a gives the distribution of ticket prices on the SOM, clearly identifying the 2 larger clusters of high-priced tickets in the lower left and upper right corner, with some additional higher-priced tickets in the upper middle and lower right corner. Fig. 2b contains the same representation with respect to the number of adults listed on the group tickets, showing a clear overlap with the distribution of higher-priced tickets, as the price of the ticket increases with group size. The distribution of the number of children (half-price tickets) is depicted in Fig. 2c, showing the tickets of groups with large number of children in the lower right corner of the SOM. Fig. 2d finally depicts the distribution of visitors nationality, with the majority of units covering tickets from predominantly Czech visitors (grey). Tickets from international groups can be found in the bottom area of the SOM (black), with German nationality groups forming a separate cluster in the lower left area (white).

Thus, by analyzing the SOM, a fast and intuitive overview of the customer groups can be obtained, and an estimate of their size is provided by the number of tickets mapped onto each area.

**V. AUTOCCLASS - PRINCIPLES AND INTERPRETATION**

*AutoClass* is an automatic classification program to extract useful information from databases [1]. It is an
approach to unsupervised classification based upon the classical mixture model, supplemented by a Bayesian method for determining the optimal classes. We emphasize that no current unsupervised classification system can produce optimal on its own. It is the interaction between domain experts and the machine searching over the model space, that generates new knowledge. Both bring unique information and abilities to the database analysis task, and each enhances the others' effectiveness.

In the AutoClass approach, class membership is expressed probabilistically rather than as logical assignment. Thus every one of n objects is considered to have a probability that it belongs to each of the possible classes. These class membership probabilities must sum up to 1 for each object, because each object must belong to some class, even though there is insufficient information to say to which class. There are no boundaries and no class assignments, only the membership probabilities [6].

In AutoClass the functional form of probability distributions is referred to as the class model. A class is defined as a particular set of parameter values and their associated model. A classification is defined as a set of classes and the probabilities of each class. The classification problem is to first choose an appropriate class model (or set of alternate models), then to search out good classifications based on these models, and finally to rate the relative quality of the alternative classifications. AutoClass generates various reports that can give a way of viewing the current classification. The "strength" of a class within a classification can usually be judged by how dramatically the highest influence value attributes in the class differ from the corresponding global attributes.

VI. AUTOCLASS EXPERIMENTAL RESULTS

We tried to find a suitable form of results presentation for the classes generated on the ticket sales data. Finally we decided to use a kind of decision tree, where nodes are attributes which mainly contribute by splitting of examples into various clusters as depicted in Fig. 3. Edges define particular values of given attributes which are typical for examples on given path. Clusters are represented as leaves in this decision tree similar to the squares in the grid from the previous SOM approach (cf. Fig. 3), with each unit again being labeled with the attributes considered most important to describe the tickets from particular cluster. Similarly, the price of the tickets in the clusters is expressed by its color, with higher-priced clusters being indicated by darker shadings.

The best clustering we got using AutoClass produced 11 clusters. The most significant attribute was price for services and order respectively (they express the same information). These attributes clearly separate one cluster with tickets sold based on orders, which are usually tickets for large groups and therefore with highest prices, as opposed to tickets sold to individual tourists at the entrance gate.

The second most important attribute for clustering was total price of the ticket separating a second cluster with expensive tickets for large groups bought with no pre-ordering. As third attribute, the nationality of visitors further split the data into two groups - tickets sold to foreigners and tickets for Czech visitors with the latter forming the vast majority of all visitors.

At the last level the date attribute (which has been transformed into days in week) did the final clustering into other 9 clusters.

VII. COMPARISON OF RESULTS

The SOM turned out to provide a quite intuitive mapping of the high-dimensional input data onto the 2-dimensional map, which allowed us to easily obtain an overview of the data characteristics. Furthermore, the SOM turned out to be very stable in terms of parameter selection, such as different sizes for the SOM or variations in learning rate parameters. However, as the SOM can only process numerical data, the transformation and normalization process was, although quite straightforward for most cases, crucial. For the interpretation of the resulting SOM we relied heavily on the LabelSOM approach to automatically extracting the most decisive attributes from the trained map. This allowed us to analyze the respective attributes in more detail, combining attribute-wise analysis with the confidence information provided by the LabelSOM approach.

The interpretation of results using the AutoClass approach turned out to be less intuitive than the classification map provided by the SOM analysis. It is necessary to go through various reports generated by AutoClass and combine the information needed to interpret clusters defined by AutoClass, though some of the examples (cases) can "lie on boundaries" between two or even more classes. However, once the decision tree is extracted it allows for a very strict and intuitive result representation, with the decisive attributes clearly standing out. Still, it has to be kept in mind, that those 'strict' classification rules are merely a result representation, or rather interpretation, as the class probabilities have to be considered.

VIII. CONCLUSIONS

We presented a comparison of two approaches to cluster analysis, applied to a real world database in conjunction with the GOAL project. Sales data of admission tickets to castles were analyzed to identify visitor clusters with respect to group sizes, nationality and high frequency periods.

Based on these results the requirements for additional information to be collected as well as additional preprocessing being required will now be analyzed. As the information from a geographic information system (GIS) will be integrated into the data warehouse (DWH) and the knowledge discovery process, more information on itinerary planning and visitor distribution shall be obtainable.
V. REFERENCES


